Constraint Based Causal discovery search on Student Achievements

Henry Valeyre
19-945-047
ETH Zürich
hvaleyre@student.ethz.ch

Javier Sanguino
21-947-791
ETH Zürich
jsanguino@student.ethz.ch

Abstract

This report aims to first find causal links using constraints based method, between parents, student motivation, classroom environment, teacher, money mental health and high school degree expectation of students, on student Achievements. In addition, this study does not aim to describe how variables affect Achievements and how to get higher student achievements but is done as a discovery of causal links. The data consists of students from the NEPS Dataset. Results found that expectations do not impact Achievements but Achievements impact the following year's expectations. In addition Achievements were time confounded. Finally, variables such as money and parents had an impact on mental health which then had a direct impact on Achievements. Furthermore, teachers and classroom quality had an impact on classroom environment which then had a direct impact on Achievements.

1 Introduction and motivation

1.1 Motivation

In Machine Learning, most of the models try to predict the outcome based on the input. These models usually try to find the best features to do a prediction but never concentrate on causal relationships. This leads to models that learn the past perfectly but that do not learn relationships that will hold in the future.

In fact, for a long time, education has greatly exploited machine learning methods. In particular, finding the relationships between two (or more) variables is really useful in this field: it can study what factors affect the learning outcome. This could lead to the design of better teaching methodologies or to build adapted learning experiences for students that have certain socio-economic background.

Nevertheless, using correlation to find relationships can be limited, specially in fields like education where hidden relationships can really affect other variables. Confounding variables can lead to reach wrong conclusions.

To really know what will happen when some variables are modified, it is needed to conduct a causal study. Causality is important to understand what is the underlying cause on observed effects.

1.2 Project description

The objective of the project was to find which important factors in the everyday life of a student would affect their performance. In order to do that, a causal study that has found the links between different aspects of students has been conducted.

More specifically, the project has been divided in two parts. In the first part, the relationship between achievements and student expectations has been studied, inspired by (Dochow and Neumeyer, 2021). We took the expectations of which high school qualification the student realistically thought they would get. We then used a PC Algorithm in order to find the causal link it has with both Achievements in German and Mathematics. Four years of study were taken in account here. Background knowledge was also added to prohibit edges to go back in time.

In the second part of the study, around 500 variables were selected as being relevant. These variables were then put together into 42 different category. Variables linked to student capacities were not taken in account. This was done because students were young and self assessment is relatively hard or linked to their own mental health or teachers. Getting the assessment of the student's capacities from parents or teachers would be biased. Thus other important factors were selected. Then PC and FCI algorithms were used to draw causal graphs.

Section 2 will discuss previous analysis and observations regarding similar variables than the one that were selected for the project. Then the constraint based causal algorithms PC and FCI will be discussed. In addition, the Dataset will be described. Finally results regarding causal discoveries will be analysed.

2 Background

The use of causality in education is recent and does not have many studies. However multiple statistical observations were made. First of all, it is important to mention that there is a constant debate whether student's expectations have an impact on their achievements. The causal analysis done on NEPS data by (Dochow and Neumeyer, 2021) showed that expectations did not have a causal relationship with achievements. The study shows as well that achievements were time confounded.

Mental health in any human being is a very important variable. However, no consensus has been reached whether achievement impacts mental health or if mental health impact achievements. The causal study of (Agnafors et al., 2021) shows that early signs of mental health problems results in poorer school performance.

Among things required to live in the world nowadays, money is one that gives you access to multiple things. Researchers are unsure of to what is the causal link between wealth of the parents and children's performances at school. A causal study done by (Duncan et al., 2011) however shows that family income has a positive impact on children's achievements.

Another important factor that impacts children is the parents. In fact they educate children and help them or not when needed. Parents can have impacts on the education of children in various manner. In this reasoning, (Helmandollar, 1992) statistically studied the impact that parents could have on their child's achievement. They found out that unlike a direct link in the achievement, parents in their implication and attitude impacted the mental state and the attitude of their child in a way that would impact its achievements at school.

Teachers are the one that unlike parents, will teach and impact the learning of the children at school. The study of (Doyle, 1977) focused on the induction of student teachers into the classroom. Their results showed that effective teachers take multiple decisions in order to adapt the class and restructure their pedagogy. This shows that teachers definitely have an impact on children and the classroom. It opens the question as to whether teachers have a direct impact on students' achievements or an indirect impact through the quality of the lesson.

Now that we mentioned the quality of the lesson and classroom, the study of (Mushtaq and Khan, 2012) showed that learning facilities, guidance of parents/teachers and student's communication skills impacted positively on the student's performance but that the family's stress impacted negatively on their performance at school.

As no analysis should be blindly followed, the project detailed in this report will use the analysis of (Dochow and Neumeyer, 2021) as a starting point. However causal observation of two variables is not enough to conclude that there are no other variables that will make them have a causal relationship or not. This is why additional variables were added to the causal discovery search on student Achievements.For this reason, and for the reasons that the variables mentioned in this section are important for the achievement of students, the scope of this project will thus, cover which variables such as students' expectations, mental health, parents, teachers, motivation, family's wealth and classroom quality, impact on achievements and how they impact on each other as well.

3 NEPS Dataset

The NEPS dataset was used according to (Blossfeld et al., 2011).

"This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, `doi:10.5157/NEPS:SC3:11.0.0`. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network."

3.1 General description of the NEPS dataset

To conduct the study, we have used the data provided by the Leibniz Institute for Educational Trajectories in Germany within the National Educational Panel Study (NEPS). This dataset is a large educational trajectories survey of students in Germany. The duration of the study is from 2009 to

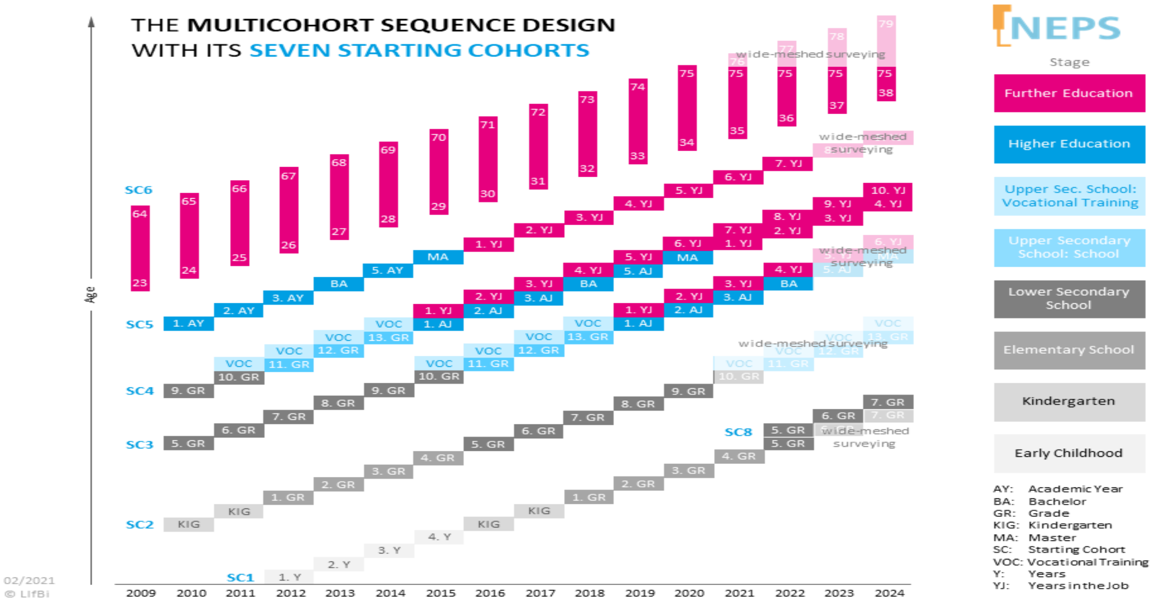

Figure 1 – Multicohort sequence design of the National Educational Panel Study
Obtained from:
https://www.neps-data.de/Project-Overview/Aims-of-the-Project

2024, hence, they have followed some students for all of their academic life.

Students in the study, their teachers, their parents, etc, are surveyed each year. The different time periods in the survey are called waves. Figure 1 shows when the data was collected. A really similar survey is done to all agents every year. Thus, an evolution in the information can be seen. This makes the dataset perfect for causal studies.

The fact that every year a lot of information is collected for each student, allows to conduct fine-grained studies with the dataset. For example, information like what visual aids their classroom has, how many hours of group projects did they have in the German class or if they live in a house where there are smokers are included in the dataset.

3.2 Specific description of data used

The dataset is organized in different cohorts, based on the age of students. For this project, we have used the data in the cohort starting grade 5 (SC3 in Image 1), which corresponds to students from fifth grade (start of secondary school) till ninth grade (first possibility to leave school). We have also selected data from wave 1 to wave 4 for the first part of the causal research, and wave 1 to 10 for the second part of the analysis(academic year 2010-2011 to academic year 2018). Wave 10 being included for the analysis of the mental health of students.

To contextualize a little bit about these years in the German education system, students (together with their parents) should have decided what their educational trajectory should be when starting fifth grade. There are different types of secondary schools chosen based on whether the students want to pursue a university degree or vocational training. Nevertheless, transferring from one to another is still possible after the first years of secondary education.

Therefore, these years are where their vocation is developed and their expectations start to be more realistic. Nevertheless, they still change over time and a more broad analysis can be made for causal methods.

In addition, their grades start to stabilize after passing from elementary education, where students are more protected, to secondary school. Therefore, their school performance is more affected by their surroundings.

These two facts: more realistic expectations and school performance were key in our decision to use this cohort.

The third cohort of the NEPS dataset had 38 tables and around 11,000 variables stored in total in all its tables. Some the information about them can be found in the Codebook: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/11-0-1/SC3_11-0-1_Codebook_en.pdf.

4 Methodology

4.1 Reproducibility

All code used can be accessed in https://github.com/fjsanguino/causal-education

4.2 Data preprocessing

To obtain the data, an NDA was signed and sent to the Leibniz Institute. That is why no data files can be found in the GitHub repository.

4.2.1 Achievements based on expectations

Once the data was downloaded, the paper by (Dochow and Neumeyer, 2021) was followed to obtain the data of achievements and expectations. The data was exported to MySQL to join tables in a easier wave

Regarding the achievements, we differentiated between the results obtained by the student in German and Math, to do a more fine-grained study. For expectations, the realistic expectations of highest school leaving qualification in waves 1,2,3 and 4 were used.

After selecting variables, they were joined together with a MySQL query using the student id and the wave. Students with one missing value for the selected waves were dropped. Furthermore students that were in a special need school were removed. This results in a dataset of 2189 students with 12 features (4 expectation waves, 4 German achievement waves and 4 Mathematics achievement waves).

4.2.2 Achievements based on more general variables

The Codebook was carefully inspected in order to pick variables that were going to be used in the project. Variables were selected taking into account the categories that we wanted to inspect for the study, its description and if it had a significant amount of data available from the selected waves. 559 variables were selected from 5 tables. They were then aggregated in 39 categories.

The output of this first preprocessing step was an Excel document with the name of the variable, its table, its description, a direct link to the page of the Codebook, waves that had data, its grouping category and if they needed to be scaled or modified in order to be added to a specific category. It is available in https://docs.google.com/spreadsheets/d/1roFaW9ptpUaahvz42P4eDOcKYxadD2hxsxv9o-GEe0E/edit#gid=0.

id	Wave	Var 1	...	Var j
Student 1	3	5		-90
Student 1	5	2		6
...				
Student 1	9	2		1
...				
Student i	3	-90		1
Student i	5	4		3

Table 1 – Example table in MySLQ Database

Afterwards, depending on the waves and the variable itself(stable parent, or changing classroom over years), a distinction between static and dynamic categories was made. It can also be seen in the Excel document.

After selecting variables, the data was exported to a MySQL database as the original downloaded data was in SPSS format. MySQL was quicker to query and merge different tables.

MySQL tables had the format shown in table 1.

Which contained the following information from selected NEPS tables:

- Dynamic data from parents, student and classes (german, math and classroom) surveys. These were obtained after merging several tables with MySQL using the student id and the waves.
- Static data from the student
- Static data from the parents
- Static data from the teacher

For the PC algorithm a preferred table was one which contained only one row per student. This is for speed and efficient analysis reasons. In the table of dynamic variables, only waves 3 and 5 were used to restrict the amount of data and they were put in two different columns. For the tables with static data, the rows of the same student with different waves were put together as one. If these static variables had values for more than one wave, the mean between them was done, but normally they had the same value in all waves. Then, all tables were merged using the student ID. All of this prepossessing was done using Python. Table 2 shows the format supposing Var 1 is dynamic and Var j is static.

id	**Var1 (w3)**	**Var1 (w5)**	**...**	**Var j**
St. 1	5	2		-90
...				
St. i	-90	-4		1

Table 2 – Objective table

The program that collapses and joins this tables is called database_to_one_id.py

After obtaining this table, there were a lot of missing values. They were filled using the Nearest Neighbors Imputation proposed by (Troyanskaya et al., 2001) using 100 groups.

Then, there were variables that shared a category but their scale was different. For example, in mental health, variable t66800l: asked if the student tended to feel depressed and variable t66800n: asked if the student showed a lot enthusiasm. One of them had to be reversed so that they had the same objective. This was done with the following equation:

$$new_{val} = max(a_i) + min(a_i) - old_{val}$$

With a_i being every element on the column that old_{val} is in. Afterwards, all values were also scaled within the same column between the min and max using a MinMaxScaler. Finally, columns were grouped into categories calculating the mean between corresponding variables. At the end, the Dataset contained 3617 students with 43 different features.

4.3 Causal study

4.3.1 PC Description

The PC algorithm is one of the oldest algorithm for Causal Study. In fact it assumes that there are no confounding variables that are not given to the algorithm. It follows the Markov condition and the faithfulness assumption , where two variables are directly causally related only if there are no subset on which they can be conditioned to be found independent.

The PC algorithm starts by forming a complete undirected graph using all variables given. Then it forms a loop that tries to find independence between variables conditioned on all other subset of variables. This means that it starts at the order 0 where every pair is conditioned on the empty set, then it goes to the order 1 where every pair of variables are conditioned on every 1 variable different from the pair and so on until every independence are found. Each conditional independence are calculated according to a given statistical procedure. A v-structure is the orientation of A to B and C to B, which exists if B was not in the set that made A and C independent. However orientation propagation can be performed where A is oriented towards B and B to C only if A and C are not adjacent. If a pair is found to be independent then the edge is removed between them. However if they are independent according to a set then variables in the set must each orient towards one of the variable in the pair. Usually the PC algorithm tries to not introduce additional v-structure when it is not needed, thus leading to orientation propagation.

4.3.2 Fast Causal Inference (FCI) Description

The FCI algorithm is a variation of the PC algorithm. Compared to the PC algorithm, FCI assumes that the world is not fully observable and that a latent confounder could exist. In this respects it tolerates confounding variables and could even find new ones. It was as well found to be asymptotically correct in the presence of latent confounders.

Different graph edges can thus be generated. If a latent confounder is found, then the algorithm will orient the edges bidirectionally. However if it is unsure it will add a round rooted arrow edge. Otherwise, the edge will be directed from A to B or non existent if it can be found conditionally independent with regard to a subset S.

4.3.3 Causal Methodology

In order to do our analysis, we used the PC and FCI algorithms from the Causal-Learn python package. For the following description, the PC and FCI algorithms used the Fisher's Z conditional independence test. This was chosen for the sole reason that not much difference was observed using other independence testing, and that fisherz belongs to the class of exact tests. Furthermore, our data in most part is categorical where fisherz performs the best, we simply transformed these categories into values. Finally, we did not have missing values on purpose, so mvfisherz was useless, and fisherz is faster than kernel based conditional testing and more precise for our specific dataset than chi/G-squared independence testing. In addition both models were used with a p-value of 0.01 for more precise results. For the PC algorithm we used uc sepset in order to orient unshielded colliders for it will give us more control to choose the priority in collider conflict solving. Prioritizing the stronger collider was our main idea, however it did not have much difference with prioritizing existing colliders, but was slower. We thus gave priority to existing colliders.

First, we started by using the PC algorithm on the German Achievement/Expectation dataset and the same on the Mathematics Achievements/Expectation dataset. Then we redid the same experiment but this time with background knowledge. The background knowledge consisted of forbidding edges to be oriented in the past. This means that achievements in t or t+ years could not impact achievements or expectations of the t- years and similarly for expectations.

Then the PC algorithm was used on the dataset with all variables. Then the PC algorithm was used again but this time with Background knowledge that dictated not only that edges could not go in the past but as well that Expectations could not impact Achievements as we previously demonstrated using the German and Mathematics Achievements.

Finally, what was done on the PC algorithm with all variables was done similarly for the FCI algorithm.

5 Results

Causal graphs obtained from the PC algorithm of German achievements and Mathematics achievements can be seen in figures 2 and 3.

The difference between without background knowledge and with background knowledge simply changes edges from undirected to directed. Background knowledge graphs show that expectations are time confounded and impact on each other. This is similar for Achievements. Finally expectations do not impact Achievements but Achievements impact Expectations of the following year.

The causal graph obtained from the PC algorithm on all variables can be seen in figures 4.

Interestingly enough, we can observe that without background knowledge, now expectations can impact achievements, but that is simply for expectation of the 5th year. Which could either show that expectations impact achievements under certain conditions, for example, all variables are observed together or later years expectations could have more importance. Another way to see it, could be that this specific year was an error.

We can as well observe that money impacts the school qualifications of parents. However we can as well see that qualifications of parents impacts on their professional qualification which in return impacts money. What is really surprising is that household money and skills of teachers are intrinsically linked. This could mean that money impacts on which teacher or institution students go to. Finally money seems to have an impact on the expectations of the students and their mental health.

Additionally it seems that how comfortable a student is at school and how curious they are, impact their motivation. But the implication of parents apparently has an implication on motivation. It can as well be observed that motivation impacts the students mental health and how parents perceive their child's mental health which shows that motivation is a confounder of mental health perception's impact on mental health of the student.

Furthermore, it can be seen that skills of a teacher impact how involved the teacher is with students, the diversity of the teaching and the teacher's critical thinking. In return, the teacher's involvement, critical thinking, as well as the diversity of teaching, teacher's experience and motivation impact the quality, diversity and positivity of classrooms. The diversity of the teaching impacts as well achievements of students. And the teacher's involvements impacts how comfortable a student is in class. In addition, the motivation of teachers seems to impact their experience.

It seems as well that the money, motivation of students, student's curiosity, implication of parents and mental health of students are connected to the pressure put on students by parents. In return it seems that parents' implication impacts the student's mental health and mental health perception, as well as the student's motivation, curiosity and comfort at school, and the resources available at home for students. Comfort is in exchange as well impacted by the amount of resources there are at home. We can thus observe some confounders here from 1st degree and 2nd degree between parents implication and comfort and curiosity. This is the same for resources and parents' implication on comfort.

Finally, we can observe that comfort at school, classroom quality and diversity, mental health and variety of teaching impacts the Achievements of students.

The causal graph obtained from the FCI algorithm on all variables with background knowledge can be seen in figures 5. The causal FCI graph without background knowledge is not displayed for

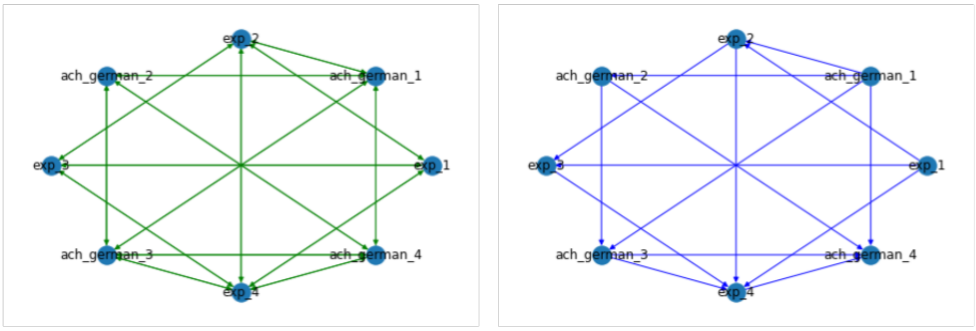

Figure 2 – Achievements and expectations causal graph for German without (left) and with (right) background knowledge

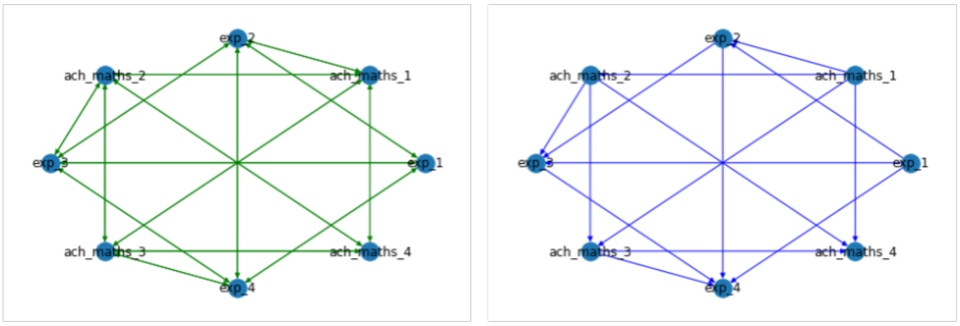

Figure 3 – Achievements and expectations causal graph for Math without (left) and with (right) background knowledge

spatial reasons.

It shows similar results than the PC algorithm however it displays that most connections are confounded by a latent variable. Interestingly enough we have seen already during the PC causal graph analysis that there already are a lot of variables that are confounded by observable ones which could be one of the reason for the appearance of many latent confounders.

A final common conclusion of PC and FCI is that immigrant background of teacher, students and parents, had no impact whatsoever on any variables.

6 Conclusion

In conclusion, we can observe that achievements are time confounded and not impacted by expectations but achievements impact expectations of the following years. Furthermore, we observed that expectations of later years could possibly impact achievements. In addition we observed that money and parents impact the mental health and curiosity of students which in exchange impacted achievements. In addition we as well observed that teachers impact diversity and quality of classrooms as well as comfort of students and diversity of teaching which are variables that directly impact Achievements. An important remark is that immigrant background of teacher, students and parents, had no impact whatsoever on any variables. Finally we observed that most of these variables are confounded by each other and could be confounded by other variables that were not observed such as student specific variables.

One of the main challenges of the project was dealing with the huge amount of data that the NEPS has. Diving into the Codebook to select variables was time-consuming. It was impossible to dedicate time to find and read bibliography of how all these variables link to the performance of a student. An improvement of the project could be doing a review of what aspects affect the learning outcome beforehand without introducing selection bias. In that sense, adding a person with an educational background to the project would be very positive.

The temporal nature of the dataset was challenging too. Dynamic data, that changed from year to year (like the grades or the classroom conditions), had to be mixed with static data, that was constant during all the years of the study (like students age or number of parents). This fact increased the time spent preprocessing the data.

To continue with the project, new analysis tools can be used as well as additional variables or different models with different parameters.

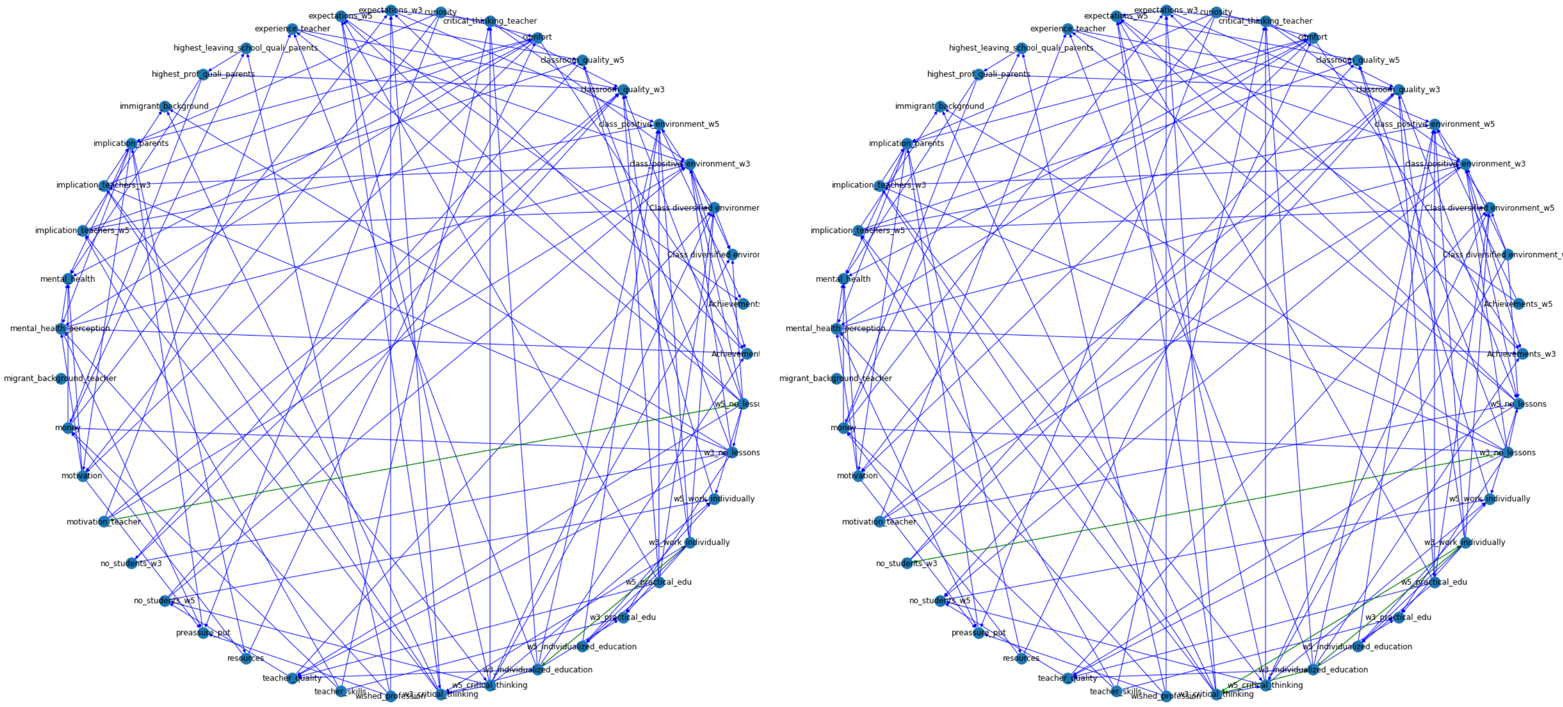

Figure 4 – PC algorithm performed on all variables without(left) and with(right) Background Knowledge

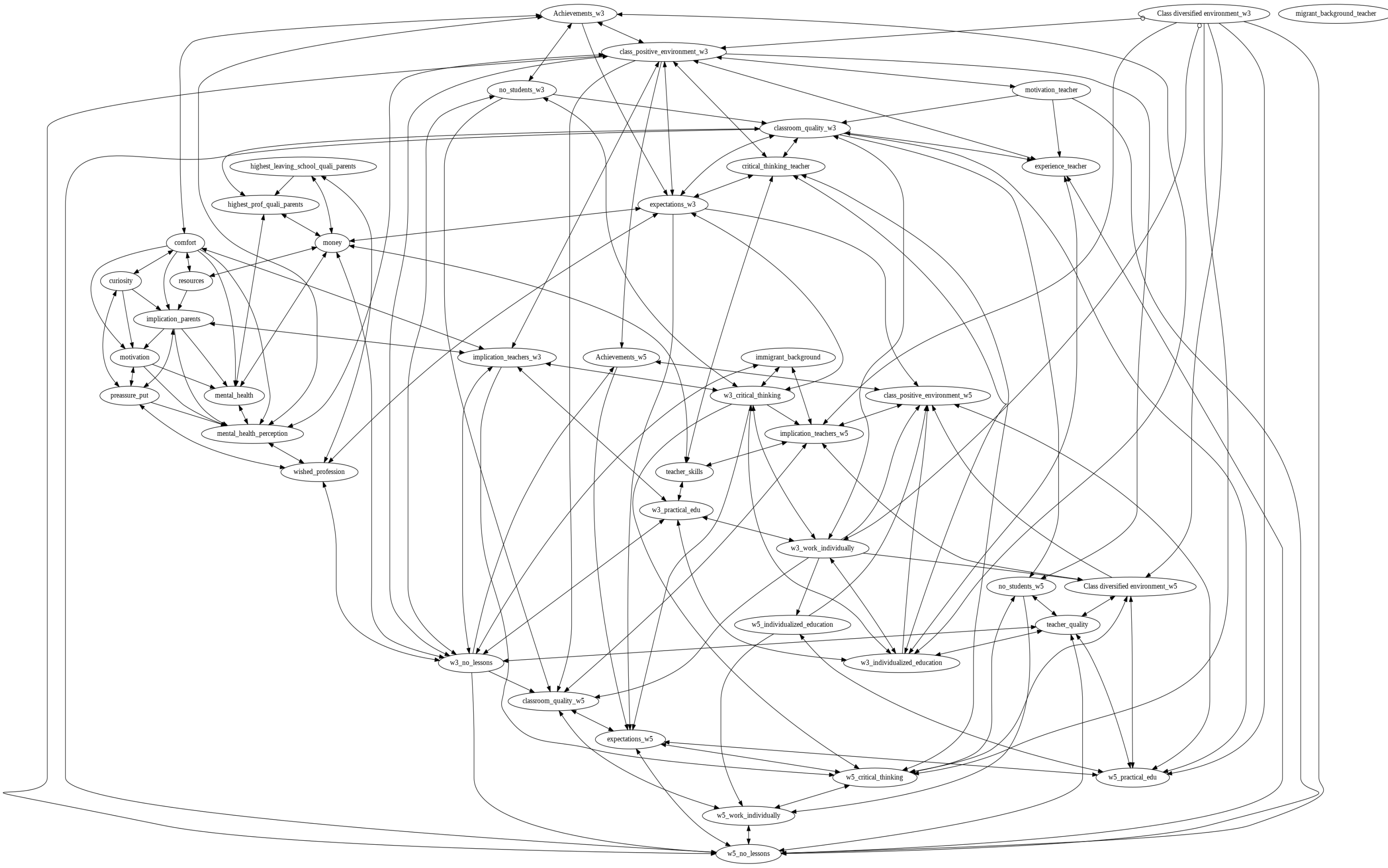

Figure 5 – Fast Causal Inference on all variables with Background Knowledge

References

Sara Agnafors, Mimmi Barmark, and Gunilla Sydsjö. 2021. Mental health and academic performance: a study on selection and causation effects from childhood to early adulthood. *Social psychiatry and psychiatric epidemiology*, 56(5):857–866.

HP Blossfeld, HG Roßbach, and J von Maurice. 2011. The german national educational panel study (neps). *Zeitschrift für Erziehungswissenschaft: Sonderheft*, 14.

Stephan Dochow and Sebastian Neumeyer. 2021. An investigation of the causal effect of educational expectations on school performance. behavioral consequences, time-stable confounding, or reciprocal causality? *Research in Social Stratification and Mobility*, 71:100579.

Walter Doyle. 1977. Learning the classroom environment: An ecological analysis. *Journal of teacher education*, 28(6):51–55.

Greg J Duncan, Pamela A Morris, and Chris Rodrigues. 2011. Does money really matter? estimating impacts of family income on young children's achievement with data from random-assignment experiments. *Developmental psychology*, 47(5):1263.

C Ben Helmandollar. 1992. *How can parents affect high school student performance by what they do at home?* Ph.D. thesis, Virginia Tech.

Irfan Mushtaq and Shabana Nawaz Khan. 2012. Factors affecting students' academic performance. *Global journal of management and business research*, 12(9):17–22.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.